文章编号: 1006-2475(2008)09-0085-03

基于 Web使用挖掘的网站优化服务研究

王 娟[1,2]

(1.厦门大学自动化系,福建 厦门 361005; 2.漳州师范学院计算机科学与工程系,福建 漳州 363000)

摘要: Web使用挖掘是通过分析上网过程所产生的数据,发现网络用户访问行为的隐含模式,以此优化网站的设计,吸引潜在的客户。本文就 Web使用挖掘技术在网站优化服务中的应用做了探讨和研究。

关键词: Web使用挖掘; 优化; 个性化
中图分类号: TP393 文献标识码: A

Research on Majorization Service of Website Based on Web Usage Mining

WANG Juan[1,2]

(1. Department of Automation, Xiamen University, Xiamen 361005, China;
2. Department of Computer Science and Engineering, Zhangzhou Normal University, Zhangzhou 363000, China)

Abstract: The Web usage mining finds out the implicit mode of network user's accessing by analyzing the data created during online, hereby optimizing the design of the Website and attracting the possible customers. This paper discusses and studies the majorization service of Website based on Web usage mining.

Keywords: Web usage mining; majorization; personalization

0 引 言

随着网络的快速发展,WWW 的访问量及复杂程度也日益提高,如何从用户对网站的使用情况中发现用户的应用需求,并以此优化网站的结构与内容、提供个性化的服务成为网站特别是商务网站提高服务竞争力的重要保障。Web使用挖掘成为解决这一问题的最有效的方法。

1 Web使用挖掘介绍

Web使用挖掘是 Web挖掘的一个组成部分,它通过挖掘用户上网过程中产生的数据,以抽取用户访问站点的浏览模式,网页的访问频率等重要信息。这些数据包括访问 Web服务器的日志、代理服务器日志、浏览器日志、用户数据、注册数据、用户会话或交易、cookies、书签数据以及任何人同 Web进行交互所产生的其他数据[1]。

目前,Web使用挖掘的研究主要有两个方向: 一般化的访问模式追踪(General Access Pattern Tracking)和个性化的使用记录追踪(Customized Usage Tracking)[2]。一般化的访问模式追踪主要用来获取用户的访问模式及趋势,个性化的使用记录追踪则是为了能够根据某个或某类用户调整网站内容的组织与显示。

基于 Web使用挖掘的网站优化服务则是结合上述两类研究,即分析各类 Web日志数据,预测用户的访问趋势以此改进 Web站点,使大部分用户感兴趣的页面更容易访问,链接更突出,网站结构更加合理;再通过分析具体用户对网站相关内容的访问记录制定出符合特定用户喜好的页面内容的组织,体现个性化服务。

2 基于 Web使用挖掘的网站优化服务模型

基于 Web使用挖掘的网站优化服务的整个过程可分为两个部分: 离线分析过程和在线处理过程。离线分析过程主要是对各种用户与网站交互的记录进行预处理和挖掘,得到有关网站结构优化和特定用户页面设置等优化模式的知识;在线处理过程则是将离线分析过程得到的知识作为输入传递给推荐机,推荐

收稿日期: 2007-09-03
作者简介: 王娟(1978-),女,山西黎城人,漳州师范学院计算机科学与工程系讲师,厦门大学自动化系硕士研究生,研究方向: 数据挖掘。

机在用户与服务器会话的过程中,通过处理输入的知识得到实时推荐系数,将其作为结果传送给服务器,在服务器端根据推荐系数优化网站,再通过会话将结构改进和符合特定用户个性的网页内容以可视化的方式在客户端浏览器上得以展现。这个服务实现的过程如图 1所示。

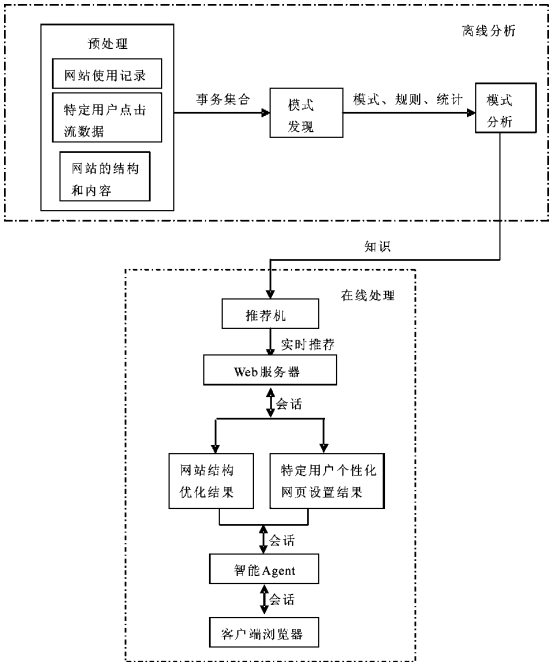

图 1 基于 Web使用挖掘的网站优化服务结构图

3 优化的具体实现过程

3.1 离线分析

在离线分析阶段主要是对 Web日志进行挖掘,具体可分为以下四个基本过程:数据收集、数据预处理、模式识别和模式分析。

3.1.1 数据收集

数据收集阶段主要是获取 Web使用挖掘所需的数据,可以从三个位置获取这部分数据:

(1)服务器端数据。通常这部分数据保存在服务器的 Web日志文件中,从这里所获得的是 Web使用挖掘的主要数据。一个实际的日志如下: 2007- 08 - 07 07: 42: 59 211. 80. 181. 71 get/ edit asp 200 该日志表示 2007年 8月 7日 7点 42分 59秒来自 IP为 211. 80. 181. 71客户端的用户以 Get方法请求服务器上的 edit asp网页,访问成功。

(2)客户端数据。这部分数据可用于识别特定的用户,可通过智能 Agent获取,具体由 Java Applet或脚本代码实现。

(3)代理端数据。这部分数据类似服务器端数据,记录客户端的请求与服务器的响应情况。

3.1.2 数据预处理

通过上述三端获得的数据并不能直接进行分析,需对它们进行预处理,这个阶段是整个挖掘过程最重要的阶段,因为数据的好坏将直接影响挖掘的结果。在该阶段对收集到的数据将进行以下几个处理过程[3]:

(1)数据清洗。去掉与挖掘无关的数据,如去掉 Http状态值为 400~ 499和 500~ 599请求网页失败的记录,以及去掉扩展名为. gif. jpeg. cgi. css的文件。

(2)用户识别。在这个阶段有两个重要的问题,一是如何识别特定的用户,二是如何识别未注册用户。解决第一个问题,可以采用 IP+ Agent机制,而对第二个问题可以采用 Web GIS[4]技术,通过 IP地址识别用户所在地理位置,考虑到可能来自同一地理位置的用户有相近的访问模式及交易方式,则在后续步骤中可用来自相同或相近地理位置挖掘出的用户访问模式等知识,预测未注册用户访问的趋势,对其提供个性化服务。

(3)会话识别。将用户一段较长时间内的访问记录进行划分,把每个划分段看成一次会话。划分的最简单方法是定义一个时间戳,根据 J Pitkow验证,用户一次访问网站的平均持续时间为 25. 5分钟,通常可以将时间戳设为 30分钟,若不同网页请求的时间差超过时间戳,则看成开始一个新的会话。

(4)事务识别。使用特定的算法将用户会话分割成更小的事务,以降低粒度适合关联规则挖掘。常用的方法有引用时长和最大前向引用。

(5)路径补充。由于缓存的存在,用户在一段时间内访问相同页面时可能是通过本地缓存,这样就不会在服务器上加载相应的日志记录,这时可采用启发性规则 + 网页拓扑结构将遗漏的请求日志加到用户会话文件中。

(6)网页内容和结构的预处理。网页结构及内容的预处理常常被忽略,但这两部分与用户的使用有着十分紧密的联系,网页如何链接取决于网页浏览方式,网站内容的创建技术又决定着网站的内容和结构,而不同用户则决定着网站主页内容的设计[5]。因此可以使用分类、聚类的方法将其转换成可用于 Web使用挖掘的格式支持挖掘。

通过上述六个步骤,可生成适合模式识别的事务集合,这组集合[6]可以表示为:

浏览页面的集合 P表示为: $P = \{p_1, p_2, \cdots, p_n\}$

用户事务集合 T表示为: $T = \{t_1, t_2, \cdots, t_s\}$

每一个事务 $t \in T$均表示为浏览页集合 P的 n维向量:

$$t = < w(p_1, t), w(p_2, t), \cdots, w(p_n, t) >$$

$$w(p_i, t) = \begin{cases} 1, & \text{if } p_i \text{ accessed by } t \\ 0, & \text{otherwise} \end{cases}$$

3.1.3 模式发现

在这一过程中主要是运用一些挖掘算法对预处理后的事务集合进行模式发现, 模式发现的主要方法有[7]:

(1)统计分析。通过 Web统计工具统计出最常访问的页面, 页面的平均访问时间, 平均访问路径的长度, 有限的错误分析等有关网站使用的基本信息, 分析用户对网络内容的关心程度。

(2)关联规则。关联规则是描述在一个事务中事件之间同时出现的规律的知识模式[2], 在优化系统中可通过关联规则分析事务集合中数据项之间的相关联系, 以此为依据调整网页链接顺序, 减少等待时间, 并可以为特定用户预设他可能关心的网页内容, 方便用户。

(3)聚类分类分析。聚类用于将事务集合中相似的数据项归并为若干个类, 在网站优化系统中, 可分两种聚类方式: 用户聚类与页聚类。用户聚类可以将网页访问习惯相近的用户归为同一类, 为他们组织网页的内容及链接; 页聚类可以将网页内容相关的组成一类, 可用于网页搜索等相关操作上。分类主要用于发展属于特定类的用户模型, 以此对用户进行分类, 为同一类的用户提供相似的服务。

(4)序列模式。在优化系统中, 可用这种方法发现用户在一段时间内对网页内容的访问序列, 以此预测用户对网站浏览趋势, 为特定的用户预设特定的内容。

(5)频繁访问组。发现用户最频繁访问的网页路径, 将该路径加载到未包含它的网页中以此优化网站结构, 方便用户。

(6)依赖建模。建立一种表示各 Web变量之间的相互依赖性的模型, 通过模型分析用户行为, 预测用户的后续动作, 增加网页预测的准确性[8]。

上述的这六种方法在网站优化系统中可综合使用, 如可以通过聚类和分类的方法将网页与用户分组, 提供相似性操作; 通过关联规则、频繁访问组、依赖建模和统计分析方法改进网站结构; 通过序列模式为用户提供个性化的服务。

3.1.4 模式分析

在模式发现过程中通过各种方法得到了一系列统计结果、规则和模型, 在模式分析阶段则是根据网站的实际应用, 通过选择和观察从中筛选出有用的模式, 将其转换成知识, 用以指导实时的网页推荐, 提供个性化服务。在该阶段可用的方法和工具有 Webviz (pitkow)系统, 该系统可以将 WWW 的访问模式可视化; Webminer系统, 它采用的是类 SQL语言的知识查询机制; 还可以使用数据仓库的联机分析处理(OLAP)方法, 发现特定的模式。

3.2 在线处理

在线处理过程中实时推荐是最重要的一个处理环节, 它将模式分析阶段的知识作为输入传给推荐机, 由推荐机根据当前用户的会话, 得到网页的推荐系数, 以此为依据优化网站结构和提供个性化服务, 通过客户端与服务器的交互, 最终传给客户端浏览器展现优化结果。其中推荐系数可表示如下:

$$Rec(S, P) = \sqrt{weight(P, C) * match(S, C)}$$

其中: S表示用户的会话集合, $S = \{s_1, s_2, \cdots, s_n\}$

P表示请求的网页, $P = \{p_1, p_2, \cdots, p_n\}$

C为用户共同的浏览模式, $C = \{w_1^c, w_2^c, \cdots, w_n^c\}$

$$\text{其中 } w_i^c = \begin{cases} weight(p_i, c), & p_i \in C \\ 0, & p_i \cup C \end{cases}$$

match(S, C)表示 C与当前用户会话的匹配度, 具体为

$$match(S, C) = \sum_k (w_k^c * s_k) / \sqrt{\sum_k (s_k)^2 * \sum_k w_k^c}$$

推荐系数大的网页则作为实时推荐的内容推荐给用户, 实现网站的优化功能。

4 结束语

如何实现网站结构优化与网站的个性化服务是当前网站建设的两大热点, 而 Web日志挖掘则是热点的研究方向。本文提出了基于 Web使用挖掘, 结合结构优化与个性化服务的网站优化体系结构的框架, 具体介绍了离线分析与在线处理两个阶段的实现过程。接下来, 还将对如何提高挖掘的效率和增强系统的安全性等方面做进一步的研究。

参考文献:

[1] Mehmed Kantardzic. 数据挖掘——概念、模型、方法和算法 [M]. 北京: 清华大学出版社, 2003. 154-163.

[2] 蒋英华. 基于 Web日志的数据挖掘 [D]. 天津大学电子与信息工程学院硕士学位论文, 2005.

[3] 刘立军, 周军, 梅红岩. Web使用挖掘的数据预处理 [J]. 计算机科学, 2007, 34(5): 200-201.

[4] 毛克彪, 等. 基于 Web GIS的电子商务数据挖掘研究 [J]. 测绘学院学报, 2003, 20(3): 180-182.

[5] 张超林, 刘丽珍, 陈俊杰. Web使用挖掘中网站结构和内容的作用 [J]. 太原理工大学学报, 2006, 37(5): 94-97.

[6] 陶剑文, 黄崇本. Web Usage Mining在网络教学中的应用研究 [J]. 情报杂志, 2006, 25(5): 73-74.

[7] 李雄飞, 李军. 数据挖掘与知识发现 [M]. 北京: 高等教育出版社, 2003.

[8] 朱明. 数据挖掘 [M]. 合肥: 中国科学技术大学出版社, 2002.